DOCUMENT RESUME

ED 430 994 TM 029 820

AUTHOR Wang, Jianjun
TITLE An Illustration of the Least Median Squares (LMS) Regression Using PROGRESS.
PUB DATE 1999-04-00
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Software; *Estimation (Mathematics); *Least Squares Statistics; Prediction; *Regression (Statistics)
IDENTIFIERS Longitudinal Study of American Youth; *Mean (Statistics); *Median (Statistics)

ABSTRACT

The least mean squares (LS) regression method produced the best linear unbiased estimates under the normal error distribution. However, many researchers have noted that the optimal condition is rarely met in real data analyses. To remedy the impact of potential data contamination, several advantages of the least median squares (LMS) regression are illustrated using a user-friendly software program, "Program for RObust reGRESSion" (PROGRESS). A public data base (the Longitudinal Study of American Youth was carefully chosen to facilitate verification of the empirical comparison between LS and LMS estimation. It is found that the LMS method results in a smaller average error of prediction and covers a larger proportion of variance in regression. In addition, it is demonstrated that even for real data with no significant outliers, the LMS estimator tended to match observations better than the simple LS fit. (Contains 2 tables and 16 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 994

RUNNING HEAD: LMS Regression

An Illustration of the Least Median Squares (LMS) Regression Using PROGRESS

Jianjun Wang
Department of Teacher Education
California State University
9001 Stockdale Highway
Bakersfield, CA 93311-1099

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Jianjun Wang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029820

Paper presented at the 1999 Annual Meeting of the American Educational Research Association, Montreal, Canada.

BEST COPY AVAILABLE

An Illustration of the Least Median Squares (LMS) Regression Using PROGRESS

Abstract

The least mean squares (LS) regression produced the best linear unbiased estimator (BLUE) under the normal error distribution. However, many researchers noted that the optimal condition was rarely met in real data analyses. To remedy impact of potential data contamination, several advantages of the least median squares (LMS) regression was illustrated in this article using a user-friendly software, Program for RObust reGRESSion (PROGRESS). A public data base was carefully chosen to facilitate verification of the empirical comparison between LS and LMS estimation. It was found that the LMS method has resulted in a smaller average error of prediction, and covered a larger proportion of variance in regression. In addition, it was demonstrated that even for real data with no significant outliers, the LMS estimator tended to match observations better than the simple LS fit.

An Illustration of the Least Median Squares (LMS) Regression Using PROGRESS

Classical regression analyses were based on the least mean squares (LS) methods which minimized the sum of residual squares in a linear regression (Casella & Berger, 1990). Weisberg (1985) noted, "One main qualification of least squares estimation is that it has been used successfully for over 150 years" (P.251). Meanwhile, many researchers expressed concerns regarding sensitivity of the LS estimators to outliers in real data analyses (e.g., Birkes & Dodge, 1993; Carroll & Ruppert, 1988; Montgomery & Peck, 1982; Rawlings, 1988). Rousseeuw and Leroy (1987) cautioned: "Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers, without careful inspection or screening" (p. vii). Although some researchers suggested that "Outlier detection procedures should be considered before any formal testing is done" (Cook & Weisberg, 1982, p. 2), they also acknowledged that "If a set of data has more than one outlier, the cases may mask each other, making finding outliers difficult" (Weisberg, 1985, p. 117).

Alternatively, Weisberg (1985) asserted, "we can think of using statistical methods that can tolerate or accommodate some proportion of bad or outlying data" (p. 116). Birkes and Dodge (1993) suggested that "The LMS [Least Median Squares] estimate is simple to describe and is very robust against outliers" (P. 207). Thus far, the LMS approach has been developed in a computer software entitled Program for RObust reGRESSion (PROGRESS), and according to Rousseeuw (1984), "The resulting estimator can resist the effect of nearly 50% of contamination in the data" (p. 871). This development was identified a frontier in statistics (Carroll & Ruppert, 1988), and as a result, PROGRESS has been integrated in the workstation package S-PLUS of Statistical Sciences (Rousseeuw & Leroy, 1987). However, few researchers in the educational research community are aware of the Least Median Squares (LMS) method. The purpose of this study is to illustrate some advantages of the LMS regression through empirical data analyses.

Literature Review

The LS method produced the best linear unbiased estimators (BLUE) under the normal error distribution (Birkes & Dodge, 1993). For real data not meeting the normality assumption, the LS fit may not be optimal. In particular, a single outlier in a data set can have profound impact on the LS estimates (Weisberg, 1985). Chatterjee and Hadi (1988) reviewed:

> Several procedures exist for the detection of a single outlier in linear regression. These procedures usually assume that there is at most one outlier in a given data set and require that the label of the outlying observation is unknown. (p. 80)

To remedy data contamination in larger proportions, robust approaches were developed to "fit a regression that does justice to the majority of the data" (Rousseeuw & Leroy, 1987, p. vii). Birkes and Dodge (1993) elaborated:

> The robustness of an estimate against heavier contamination is measured by it breakdown point, which is the least proportion of outliers that can occur in a sample without entailing the possibility of arbitrarily large bias. (P. 207)

In the LS estimation, "A single point far removed from the other data points can have almost as much influence on the regression results as all other points combined" (Rawlings, 1988, p. 241). Thus, the LS estimate can be seriously disturbed by data contamination because of the zero breakdown point in LS modeling (Rousseeuw & Leroy, 1987).

A higher breakdown point in robust regression was an important feature ameliorating weakness in outlier diagnostics. Cook and Weisberg (1982) acknowledged, "the use of robust methods does not abrogate the usefulness of diagnostics in general, although it may render certain of them unnecessary" (p. 2). According to Rousseeuw and Leroy (1987),

> Diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points, after which these outliers can be removed or corrected, followed by an LS analysis on the remaining cases. When there is only a single outlier, some of these methods work quite well by looking at the effect of deleting one point at a time. Unfortunately, it is much more difficult to diagnose outliers when are several of them. (p. 8)

For the multiple outlier cases, Rousseeuw and Croux (1993) noted, "The median has a

breakdown point of 50% (which is the highest possible), because the estimate remains bounded when fewer than 50% of the data points are replaced by arbitrary number" (p. 1273). Birkes and Dodge (1993) concurred:

> The maximum possible breakdown point is 50%. This is achieved by the least-median-of-squares (LMS) estimate, which is the estimate that minimizes the median of the squared residuals e_i^2 (or, equivalently, minimizes the median of the absolute residuals $|e_i|$). (p. 207)

The evolution from the LS to LMS estimators depends on development of the modern computing technology. Rousseeuw and Leroy (1987) recollected:

> At the time of its [LS estimator] invention (around 1800) there were no computers, and the fact that the LS estimator could be computed explicitly from the data (by means of some matrix algebra) made it the only feasible approach. Even now, most statistical packages still use the same technique because of tradition and computation speed. (p. 2)

Meanwhile, Rawlings (1988) observed:

> The method of ordinary least squares gives equal weight to every observation. However, every observation does not have equal impact on the various least squares results. (p. 241)

Investigation of the unequal data weight can be dated back to Bernoulli's (1777) article. Nonetheless, Rousseeuw and Leroy (1987) pointed out, "Without the aid of a computer, it would never have been possible to calculate high-breakdown regression estimates" (p. 29).

Built on the personal computer and mainframe interfaces, the PROGRESS software was an efficient tool for the LMS regression, and has been made "available for everyday statistical practice" (Rousseeuw & Leroy, 1987, p. ix). Rousseeuw and Leroy (1987) added:

> We advocate the least median of squares method (Rousseeuw 1984) because it appeals to the intuition and is easy to use. No background knowledge or choice of tuning constants are needed: You just enter the data and interpret the results. It is hoped that robust methods of this type will be incorporated into major statistical packages, which would make them easily accessible. (Rousseeuw & Leroy, 1987, p. viii).

The software user manual was published by the John Wiley & Sons company in its probability and mathematical statistics book series (Rousseeuw & Leroy, 1987). The latest upgrading was made in 1996, and both LS and LMS estimates were included in the PROGRESS printout.

Because of the wide dissemination, an illustration of the LMS regression may help enrich educational statistics methods with the latest software development. To involve more researchers evaluating the LMS regression, public data have been carefully chosen in this study to facilitate the empirical result verification.

Data Selection

The National Center for Education Statistics (NCES) is the federal agency in charge of collecting the national data on education. In the mid 1990s, a guideline was developed by the NCES (1996) requiring user licenses to access the restricted national data bases. Among the license requirement is an Attorney General's signature in each state. Consequently, most researchers with little connection at the state level cannot access the restricted data bases at NCES.

On the other hand, the National Science Foundation funded the Longitudinal Study of American Youth (LSAY) project during 1987-1992. The LSAY data are distributed by the Chicago Academy of Science with no license restriction. Up to the mid 1997, the project was cited in 22 articles in the ERIC data base, and a training session for using the LSAY data was offered at the 1997 annual meeting of the American Educational Research Association (AERA). To facilitate the empirical result reconfirmation, the LSAY data were employed in this study to illustrate the use of PROGRESS in the LMS regression.

Methods

Rousseeuw and Zomeren (1990) observed, "Outliers in a multivariate point cloud can be hard to detect, especially when the dimension p exceeds 2, because then we can no longer rely on visual perception" (P. 633). To simplify the illustration, two variables were chosen from the LSAY principle data file, one measuring the school enrollment (LSAY variable name: EK2A) and the other assessing the total number of grade levels in a school (LSAY variable name:

EK1A). In a real school setting, no enrollment, no school grade levels. Thus, the relation can be modeled in a linear equation with no fixed effect of intercept:

$$EK2A = \beta\,(EK1A) + \varepsilon \qquad (1)$$

where ε is the error term and β can be estimated through either LS or LMS regression.

The PROGRESS software was used to calculate the LS and LMS regression coefficients. The model comparison was based on the mean residual differences between the LS and LMS estimates. The pairwise t test was employed to further examine the real data deviation from the fit of LS and LMS models. The coefficient of determination (R^2) was also computed for each model to assess the overlap of variability between the independent and dependent variables.

Results

The results of LS and LMS estimations were assembled in Table 1. The LS estimates have been double-checked by the PROGRESS and SAS printout to ensure proper computing in the empirical data analyses.

Table 1 inserted around here

Inspection of Table 1 indicated different regression coefficients (β_{EK1A}) between the LS and LMS methods. The coefficient of determination revealed that a larger proportion ($R^2 = .84$) of the enrollment (EK2A) variation has been accounted for by the LMS prediction.

The t test results were presented in Table 2 to reflect the deviation between observed and predicted values of EK2A.

Table 2 inserted around here

Differences in the mean residual indicated that the LMS fit had a much smaller average

deviation from the observed enrollment. At $\alpha = .05$, the t test exhibited that the regression residuals for the LMS model were insignificantly different from zero. However, for the LS model, the residual was statistically significant ($p = .038$). Thus, the LMS model seemed more admissible according to the empirical data analyses.

Discussions

Researchers found that most real data did not meet the normality assumption to optimize the LS estimators (Cook & Weisberg, 1982). Consequently, diagnostic approaches attracted attention of most data analysts. McGinnis (1991) reviewed six diagnostic procedures, and recommended the use of Cook's D measure to detect outliers. But the Cook's D, like other options in SPSS or SAS, was based on the LS fitting (Carroll & Ruppert, 1988). Rousseeuw and Leroy (1987) pointed out that the LS reference may not expose outliers in many circumstances.

Similarly, in the BMDP software, the Mahalanobis Distance was employed to identify outliers. Stevens (1992) advocated:

> Fortunately, however, there is a statistic (called Mahalanobis Distance) which has an approximate chi-square distribution for large N, which can be used to detect multivariate outliers of any type. (P. 17-18)

Rousseeuw and Zomeren (1990) cautioned that "It is well known that this approach [the Mahalanobis Distance method] suffers from the masking effect, by which multiple outliers do not necessarily have a large MD_i [Mahalanobis Distance]" (P. 633).

With the highest possible breakdown point, the LMS estimator was insensitive to the impact of a few outliers. On the contrary, "outliers are far away from the robust fit and hence can be detected by their large residuals from it" (Rousseeuw & Leroy, 1987, p. vii). Thus, the LMS method can be employed for two purposes: identifying outliers and constructing robust regressions. In the example illustrated in this article, no significant differences were found between the LMS prediction and real observations. Despite the lack of significant outliers, the LMS method still resulted in a better model than the LS approach, covering larger variability in the regressional analysis ($R^2 = .84$).

References

Bernoulli, D. (1777). The most probable choice between several discrepant observations and the formation of the most likely induction. In C. G. Allen (1961), Biometrika, 41, 3-13.

Birkes, D. & Dodge, Y. (1993). Alternative methods of regression. New York, NY: Wiley.

Carroll, R. J. & Ruppert, D. (1988). Transformation and weighting in regression. New York, NY: Chapman and Hall.

Casella, G and Berger, R. L. (1990). Statistical inference. Pacific Grove, CA: Brooks/Cole.

Chatterjee, S. & Hadi, A. (1988). Sensitivity analysis in linear regression. New York, NY: Wiley.

Cook, R. & Weisberg, S. (1982). Residuals and influence in regression. New York, NY: Chapman & Hall.

McGinnis, J. (1991, April). A comparison of six different diagnostic procedures used to check raw quantitative data for outliers in a generic science education study. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Lake Geneva, WI.

Montgomery, D. & Peck, E. (1982). Introduction to linear regression analysis. New York, NY: Wiley.

NCES (1996). Restricted-use data procedures manual (NCES 96-860). Washington, DC: U.S. Department of Education.

Rawlings, J. O. (1988). Applied regression analysis: A research tool. Pacific Grove, CA: Wadsworth.

Rousseeuw, P. J. & Leroy, A. M. (1987). Robust regression and outlier detection. New York, NY: Wiley.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79 (388), 871-880.

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88 (424), 1273-1283.

Rousseeuw, P. J. & Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85 (411), 633-639.

Stevens, J. (1992). Applied multivariate statistics for the social sciences (2nd ed.). Hillsdale, NJ: Lawrence Erkbaum.

Weisberg, S. (1985). Applied linear regression. New York, NY: Wiley.

Table 1

A comparison between the LS and LMS estimates

Method	β_{EK1A}	Model	R^2
LS	212.20	EK2A = 212.20*EK1A	.64
LMS	236.67	EK2A = 236.67*EK1A	.84

Table 2

Test of differences between predicted and observed outcomes

Model	N	Mean Residual	Standard Deviation	T test
LS	86	-141.68	623.18	t (85) = -2.11, p = .038
LMS	86	- 50.64	644.92	t (85) = -0.73, p = .469